Segment Together: A Versatile Paradigm for Semi-Supervised Medical Image Segmentation

Qingjie Zeng, Yutong Xie, Zilin Lu, Mengkang Lu, Yicheng Wu, and Yong Xia, *Member, IEEE*

***Abstract*— The scarcity of annotations has become a significant obstacle in training powerful deep-learning models for medical image segmentation, limiting their clinical application. To overcome this, semi-supervised learning that leverages abundant unlabeled data is highly desirable to enhance model training. However, most existing works still focus on specific medical tasks and underestimate the potential of learning across diverse tasks and datasets. In this paper, we propose a Versatile Semi-supervised framework (VerSemi) to present a new perspective that integrates various SSL tasks into a unified model with an extensive label space, exploiting more unlabeled data for semi-supervised medical image segmentation. Specifically, we introduce a dynamic task-prompted design to segment various targets from different datasets. Next, this unified model is used to identify the foreground regions from all labeled data, capturing cross-dataset semantics. Particularly, we create a synthetic task with a CutMix strategy to augment foreground targets within the expanded label space. To effectively utilize unlabeled data, we introduce a consistency constraint that aligns aggregated predictions from various tasks with those from the synthetic task, further guiding the model to accurately segment foreground regions during training. We evaluated our VerSemi framework against seven established SSL methods on four public benchmarking datasets. Our results suggest that VerSemi consistently outperforms all competing methods, beating the second-best method with a 2.69% average Dice gain on four datasets and setting a new state of the art for semi-supervised medical image segmentation. Code is available at https://github.com/maxwell0027/VerSemi**

Received 8 December 2024; revised 15 March 2025; accepted 24 March 2025. Date of publication 31 March 2025; date of current version 2 July 2025. This study was supported in part by grants from National Key R&D Program of China (2022YFC2009903/2022YFC2009900), in part by the National Natural Science Foundation of China under Grant 62171377 and Grant 92470101, and in part by the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University. Recommended by Associate Editor Dr. D. Mahapatra. *(Corresponding authors: Yutong Xie; Yong Xia.)*

Qingjie Zeng, Zilin Lu, Mengkang Lu, and Yong Xia are with the National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: qjzeng@mail.nwpu.edu.cn; luzl@mail.nwpu.edu.cn; lmk@mail.nwpu.edu.cn; yxia@nwpu.edu.cn).

Yutong Xie is with Australian Institute for Machine Learning, The University of Adelaide, Adelaide, SA 5000, Australia (e-mail: yutong.xie678@gmail.com).

Yicheng Wu is with the Department of Data Science and AI, Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia (e-mail: yicheng.wu@monash.edu).

Digital Object Identifier 10.1109/TMI.2025.3556310

***Index Terms*— Semi-supervised learning, medical image segmentation, unified learning.**

I. Introduction

MEDICAL image segmentation remains a formidable challenge, despite significant advances in the field [1], [2], [3], [4], [5]. The scarcity of voxel-level annotations has led to the adoption of semi-supervised learning (SSL) strategies, which effectively utilize a combination of limited labeled and extensive unlabeled data in medical image segmentation tasks.

Generally, there are two popular SSL paradigms, *i.e.*, pseudo-labeling [7] and consistency regularization [8], [9]. The former aims to generate reliable pseudo-labels for re-training, for instance, by applying an adaptive threshold to exclude unreliable predictions [10]. The latter, on the other hand, seeks consistent predictions across various augmentations of the same input [11], [12], [13], [14]. Despite their widespread use, these methods are typically confined to a single task with a consistent label space for both labeled and unlabeled data. This limitation often leads to discrepancies in the learned distributions between the two data types [15], [16], resulting in suboptimal generalizability and performance of SSL techniques. Additionally, a significant portion of unlabeled data may be underutilized when it does not align perfectly with the predefined task label space.

The advent of universal models has drawn attention for their adaptability and capacity to handle diverse tasks. These models are trained on multi-domain and/or multi-modality data for multi-task through two distinct approaches. One involves pre-training on task-agnostic unlabeled data via self-supervised learning, followed by fine-tuning on task-specific labeled data for individual downstream tasks [17], [18], [19]. The other approach trains a model jointly using multiple task-specific data in a supervised fashion [6], [20], [21]. Demonstrated by their exceptional performance across a range of tasks in computer vision [22], [23] and medical image analysis [18], [21], [24], universal models unveil the significance of integrating data and tasks to improve representation learning. This insight motivates the fusion of multiple datasets and tasks within an SSL framework, promising not only to leverage a more extensive corpus of labeled and unlabeled data, thereby bolstering the supervised component of SSL, but also to significantly improve the model's generalization ability, thereby extending its applicability across diverse domains.

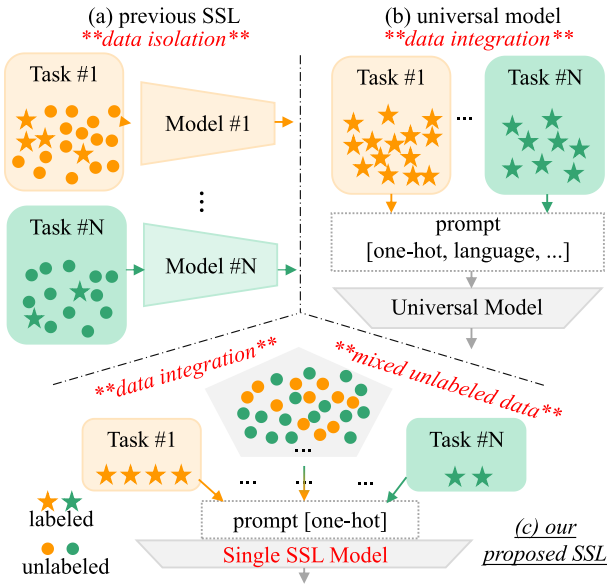

Fig. 1. Comparative illustration between (a) traditional SSL methods, (b) universal models, and (c) our proposed VerSemi framework. Traditional SSL methods typically train each model separately, overlooking the potential benefits of integrating data from multiple tasks. Universal models, such as DoDnet [6], utilize a variety of task prompts to assimilate knowledge from multiple tasks in a supervised manner but often struggle with unlabeled data lacking task-specific information. Our VerSemi framework, however, is designed to handle multiple tasks simultaneously and to learn from unlabeled data without requiring specific task details.

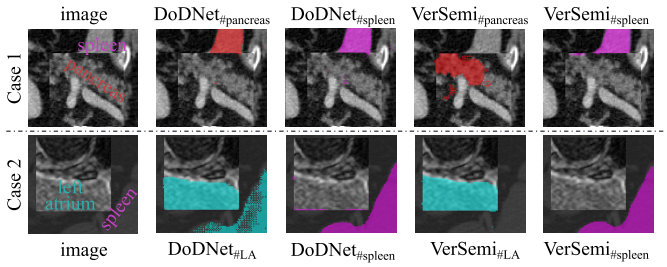

Fig. 2. Depiction of prompt-weakening phenomenon. In Case 1, DoDNet fails to correctly identify the pancreatic region, instead recognizing spleen voxels irrespective of the prompt for pancreas or spleen. In Case 2, DoDNet mistakenly highlights the spleen region when prompted for the left atrium. This issue is attributed to the shared representations learned across heterogeneous tasks. Our VerSemi framework addresses this by introducing an auxiliary constraint.

In this paper, we propose a novel **Ver**satile **Semi**-supervised framework (VerSemi), which revolutionizes traditional SSL paradigms.

First, VerSemi overcomes the limitations of task-specific learning by integrating multiple objectives into a unified framework. It adeptly establishes a comprehensive label space by amalgamating relevant task labels and simultaneously executes multiple tasks with the assistance of a task-prompted dynamic head.

Second, given that prompt-driven models necessitate task-specific details to generate prompts (*e.g.*, one-hot task code and language description) during training, an issue arises when unlabeled data remains unexploited if associated task information is unavailable (see Fig. 1). To tackle this issue, VerSemi initially constructs a synthetic task using CutMix [25] on labeled data. The data in the synthetic task cover a diverse range of foreground targets in the expanded label space. Through joint training with the synthetic task, VerSemi acquires the capability to recognize and segment all potential foreground regions. This proficiency allows VerSemi to simplify the learning from unlabeled data by obviating the need for task-specific details. This is achieved by ensuring the consistency between combined predictions from related tasks and synthetic ones.

Third, we identify that prompts may be less effective with limited annotations, as models might not recognize the object indicted by a specific prompt (see Fig.2). To address this, VerSemi utilizes an auxiliary constraint as a regularizer to enhance its controllability when meeting task-specific prompts. Notably, existing SSL methods cannot directly undertake task-agnostic learning from unlabeled data, as they typically require either a teacher model or extra sub-networks for supervision.

Our contributions are three-fold.

- We propose VerSemi, an innovative framework that excels in integrating various related SSL tasks into a unified framework, challenging the conventional task-specific SSL approaches.
- We achieve task-agnostic learning from unlabeled data through the innovation of a "synthetic task", fostering a unified foreground segmentation capability. This capability, when used as a constraint, allows for the exploration of unlabeled data without task-specific information.
- Extensive experiments on four public datasets confirm VerSemi's superiority, demonstrating remarkable improvements over both task-specific SSL models (*e.g.*, BCP [15], CauSSL [26]) and task-unified models (*e.g.*, Uni-BCP, Uni-CauSSL).

II. Related Work

A. SSL Methods

SSL has emerged as a solution to alleviate the labor-intensive process of data annotation by utilizing a mix of scarce labeled data and abundant unlabeled data [2], [27], [28], [29], [30], [31]. Many studies have sought to effectively extract information from unlabeled data. For instance, UPS [32] was designed to reduce unreliable pseudo-labels by calibrating models with uncertainty. PEFAT [7] analyzed the probability distribution of pseudo-labeled data and proposed a novel selection criterion based on loss distribution. SoftMatch [33] and FreeMatch [34] attempted to address the quantity-quality trade-off issue with adaptive thresholding. These SSL frameworks, however, are predominantly focused on learning within single tasks, which leads to the question of their scalability to heterogeneous tasks. Additionally, when learning tasks separately, there is often only marginal improvement due to the insufficient representation derived from the limited labels in each dataset.

To overcome these challenges, we propose VerSemi, a unified SSL model that learns from an integrated dataset, to enable

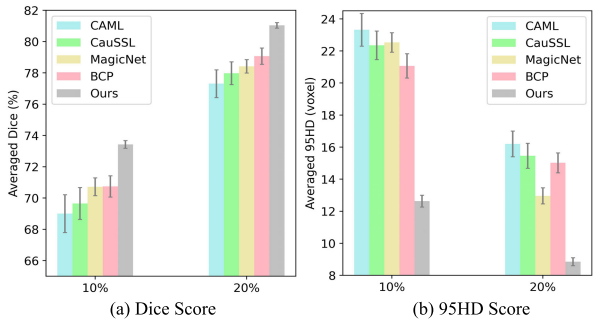

Fig. 3. Averaged (a) Dice score and (b) 95HD score of five SSL methods across four tasks. Notably, VerSemi outperforms the second-best method, BCP [15], by margins of 2.69% and 2.04% in average Dice and 8.92 and 6.56 voxels in average 95HD, with 10% and 20% labeled data, respectively. These results indicate VerSemi's ability to provide a comprehensive and versatile solution, with enhanced performance due to the effective use of both labeled and unlabeled data. This underscores the importance of transcending task-specific SSL limitations by integrating multiple objectives within a single SSL framework.

the concurrent learning of multiple related SSL tasks. Moreover, to our knowledge, existing SSL methods have been evaluated only within the context of learning tasks separately. In contrast, VerSemi is the first framework to demonstrate the capability of scaling to multiple SSL tasks simultaneously. VerSemi offers several advantages, including: (1) high performance (see Fig. 3) and efficiency; (2) strong transferability to an unseen test set (see V-G); and (3) greater versatility. Unlike existing methods that either require multiple networks for cross-supervision or rely on additional teacher models for knowledge transfer, VerSemi learns from unlabeled data using a single network without the need for task-specific information.

B. Representation Learning With Integrated Data

To improve model performance and representation ability, some researchers have proposed learning a unified model capable of executing multiple tasks simultaneously, rather than training task-specific model separately [35], [36], [37], [38], [39], [40]. DoDNet [6], for instance, compiled an abdominal dataset from seven partially labeled datasets for model training and achieved superior average results compared to the models trained on individual datasets. CLIP-Driven Universal Model [24] further incorporated CLIP embedding [41] to assist the model in capturing anatomical relationships between various tumors and organs. UniSeg [21] utilized different imaging modalities, such as CT, MRI, and PET, and outperformed the models trained on a single modality. These studies highlight the significance of robust data integration, advocating for the maximal utilization of available data. While the majority of them focus on either fully supervised learning [21], [24] or self-supervised learning [18], few have explored the simultaneous use of labeled and unlabeled data from different tasks. Our proposed VerSemi represents an innovative attempt in this direction.

Furthermore, existing universal models (*e.g.*, DoDNet [6], CLIP-Driven Universal Model [24], and UniSeg [21]) concentrate on designing appropriate task prompts to learn from heterogeneous tasks in a supervised manner. However, VerSemi is dedicated to the design of task-agnostic learning from unlabeled data, a feat not yet accomplished by current universal models, which depend on task-specific information to generate task prompts. Moreover, VerSemi tackles the critical issue of task prompt weakening, an aspect overlooked by these models (see Fig. 2). By addressing this challenge, VerSemi substantially improves the controllability of task prompts when dealing with various related tasks.

III. Method

The proposed VerSemi integrates diverse semi-supervised segmentation tasks from various datasets into a unified framework. It utilizes a task-prompted dynamic head to handle a range of assignments. To facilitate the task-agnostic learning from unlabeled data, we construct a synthetic task (Task#5) by applying CutMix [25] to labeled data across relevant tasks (Task# $\sim$ Task#4), resulting in a mixed dataset. This synthetic task serves as a guide for VerSemi to discern all potential targets. By leveraging this capability, unlabeled information is obtained by aligning the aggregated predictions (from Task#1 to Task#4) with the predictions generated by the synthetic task. This alignment is achieved by feeding mixed unlabeled data into VerSemi.

Therefore, task specifics (which task the data belongs to) are not needed for the design of prompt when confronted with unlabeled data. For both labeled and unlabeled data mining, data from different tasks are randomly sampled. The VerSemi pipeline is depicted in Fig. 4, we now delve into its details.

A. Dynamic Convolution With Task Prompt

Although deep learning models have made significant progress, a single model with fixed convolutional kernels often performs sub-optimally when handling multiple segmentation tasks simultaneously [6], [42]. A multi-head architecture is commonly employed to enhance performance, but it incurs substantial computational costs as the number of tasks increases, limiting its applicability in scenarios with multiple tasks. To reduce the computational burden associated with multiple heads, we utilize dynamic filter generation to construct the segmentation head. This approach enables adaptive processing of different tasks using task-specific prompts without introducing extra computational expenses. The filter generation process is defined as:

$$\begin{aligned} w_k &= \psi(GAP(Embedding), [Prompt_{\#k}]; \theta_\psi) \\ \mathcal{P}_k &= SoftMax(f_D(Embedding) * w_k), \end{aligned} \tag{1}$$

where ψ is a convolutional layer with parameters θ_ψ, designed to dynamically generate filter parameters w_k for the segmentation head when processing Task#k based on $[Prompt_{\#k}]$. The prompt, encoded in a one-hot format, is concatenated with a globally averaged feature embedding before being input into ψ. The symbol $\mathcal{P}k$ represents the prediction for Task#k, f_D is the decoder, and $*$ denotes convolution. GAP represents global average pooling. By using $[Prompt_{\#k}]$, VerSemi effectively identifies the current task and adjusts its kernels accordingly.

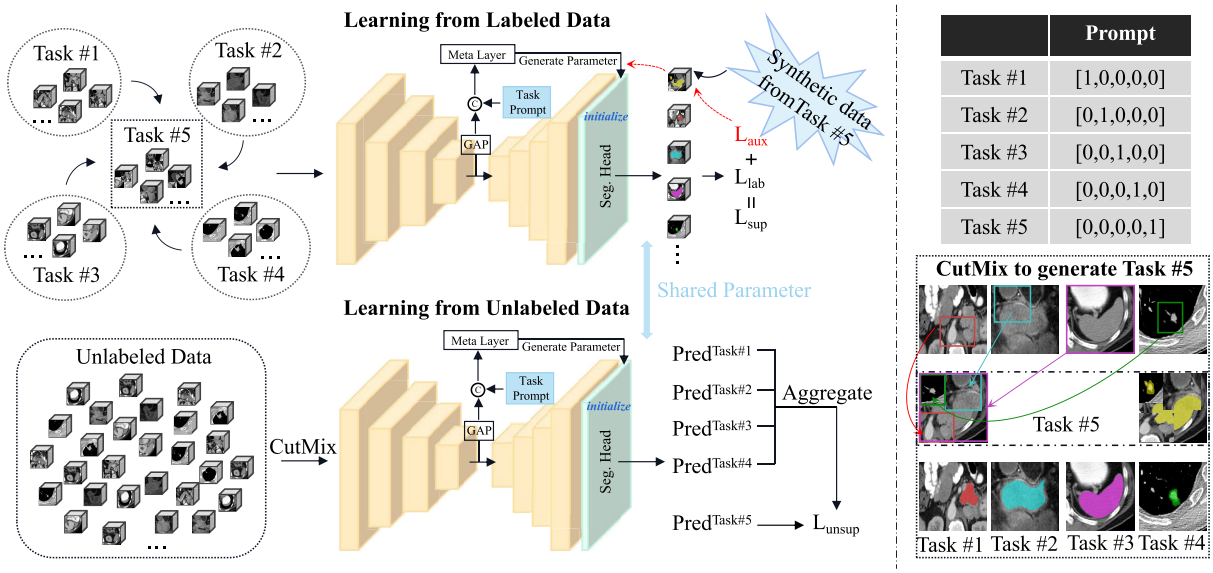

Fig. 4. Architecture of our proposed VerSemi, featuring a task-prompted dynamic head for handling various tasks concurrently and an auxiliary constraint $\mathcal{L}_{\text{aux}}$ to augment the reliability of associated task prompts. During labeled data learning, a synthetic task (Task#5) is constructed to segment all foreground regions. During unlabeled data learning, the model aligns the combined predictions, elicited by prompts from Task#1 to Task#4, with those from Task#5, thus enabling task-agnostic learning and showcasing VerSemi's adaptability. "Meta Layer" refers to a convolutional layer responsible for dynamically predicting kernel parameters for the segmentation head, and "Generate Parameter" indicates the process of using these predicted parameters to initialize the segmentation head.

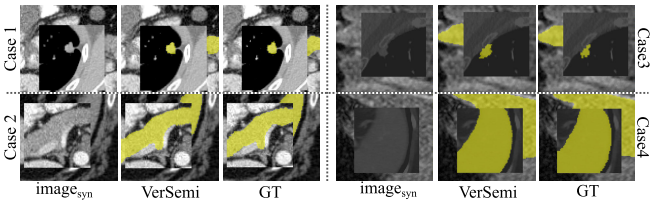

Fig. 5. Predictions made by VerSemi when processing synthetic data. Case 1 and Case 3: CutMix between left atrium and lung tumor. Case 2: CutMix between pancreas and spleen. Case 4: CutMix between spleen and left atrium. Precise foreground can be discerned when prompted by Task#5.

B. Task-Aware Labeled Data Learning

Our study incorporates four relevant tasks: the segmentation of the pancreas, left atrium, spleen, and lung tumor, designated as **Task#1** $\sim$ **Task#4**. We also introduce a synthesized **Task#5** as follows.

Generation of Task#5. As depicted in the bottom-right of Fig. 4, we construct a synthetic task (Task#5) using labeled data from the four pertinent tasks (Task#1 $\sim$ Task#4). Task#5 is designed to assist VerSemi in task-agnostic learning from unlabeled data and guide the model to segment all foreground regions in mixed data scenarios (see Fig. 5). The generation of data for Task#5 is formulated as:

$$\begin{aligned}\mathcal{X}^l_{syn(i,j)} &= \mathcal{X}^l_i \odot \mathcal{M} + \mathcal{X}^l_j \odot (1-\mathcal{M}) \\ \mathcal{Y}^l_{syn(i,j)} &= \mathcal{Y}^l_i \odot \mathcal{M} + \mathcal{Y}^l_j \odot (1-\mathcal{M}),\end{aligned} \tag{2}$$

where $\mathcal{X}^l_{syn(i,j)}$ and $\mathcal{Y}^l_{syn(i,j)}$ are the synthetic images and labels for Task#5. $\mathcal{M}$ is a mask with 30% $\sim$ 70% random masked regions. The symbol $\odot$ is element-wise multiplication. Note that $\mathcal{Y}^l_i$ and $\mathcal{Y}^l_j$ are binary masks, $\mathcal{X}^l_i$ and $\mathcal{X}^l_j$ are images from the i-th and j-th task, thus $\mathcal{X}^l_{syn(i,j)}$ can be regarded as mixed data that contain various targets and background. The Dice loss and cross-entropy loss are applied to the labeled data learning process, including Task#5, and are defined as:

$$\begin{aligned}\mathcal{L}_{lab} = Dice(\mathcal{F}(\mathcal{X}^l_k, [Prompt_{\#k}]; \Theta), \mathcal{Y}^l_k) + \\ CE(\mathcal{F}(\mathcal{X}^l_k, [Prompt_{\#k}]; \Theta), \mathcal{Y}^l_k),\end{aligned} \tag{3}$$

where $\mathcal{L}_{lab}$ is the supervised loss on labeled data. For simplicity, we use $\mathcal{F}(\cdot;\Theta)$ to denote the entire network with parameters Θ, which includes the operations from Eq.1. Benefited from Task#5, VerSemi has a semantic perception of all segmentation tasks.

Enhancing the controllability of task prompt with $\mathcal{L}_{aux}$. As observed in Fig. 2, there exists a weakening effect on the task prompt when employing a task-prompted dynamic head. Models like DoDNet [6] may occasionally fail to identify the intended task even under the control of a task-specific prompt, likely due to overlapping semantic information across segmentation tasks. To strengthen the uniqueness of the task prompt, we introduce an auxiliary constraint $\mathcal{L}_{aux}$, which is formulated as:

$$\begin{aligned}\mathcal{L}_{aux} = \sum_{k\in(i,j)} Dice(\mathcal{F}(\mathcal{X}^l_{syn(i,j)}, [Prompt_{\#k}]; \Theta), \mathcal{Y}^l_k) \\ + CE(\mathcal{F}(\mathcal{X}^l_{syn(i,j)}, [Prompt_{\#k}]; \Theta), \mathcal{Y}^l_k),\end{aligned} \tag{4}$$

where $\mathcal{X}^l_{syn(i,j)}$ represents a synthesized image containing targets from both the i-th and j-th tasks. When $k=i$ (or j), the model is constrained to focus solely on classifying the i-th (or j-th) target, where the supervision label $\mathcal{Y}^l_k$ is adapted to $\mathcal{Y}^l_{i(j)}$. In other words, this formulation ensures that the model selectively attends to the task it is prompted for, even when dealing with mixed-task data. When i is equal to j, $\mathcal{L}_{aux}$ transforms into a standard supervised loss in conjunction with

the CutMix augmentation within the i-th task. To this end, The supervised loss $\mathcal{L}_{sup}$ is thus written as:

$$\mathcal{L}_{sup} = \mathcal{L}_{lab} + \mathcal{L}_{aux}. \tag{5}$$

Consequently, VerSemi introduces the synthetic Task#5, facilitating task-agnostic learning from unlabeled data and enhancing the capability of task prompt with the help of $\mathcal{L}_{aux}$.

C. Task-Agnostic Unlabeled Data Learning

Considering task-specific information is required for prompt-driven model to generate prompt, we place our VerSemi under a more challenging SSL context, in which unlabeled task-specific information is unavailable. We detail this process as follows. First, CutMix [25] is applied to all unlabeled data, resulting in inputs that contain objects from various tasks. The prediction generated with the **Task#5 Prompt** is then forced to align with the aggregated prediction from **Task#1 Prompt** $\sim$ **Task#4 Prompt** (see Fig. 8). This aggregated prediction can be regarded as a combination of pseudo-masks for each task, while the prediction prompted by Task#5 serves as a universal pseudo-mask for all tasks. Therefore, these two predictions should be identical. This operation is termed as self-consistency, since no extra decoder or teacher model is required for supervision. The entire process can be written as:

$$\begin{aligned} \mathcal{X}^u_{syn(i,j)} &= \mathcal{X}^u_i \odot \mathcal{M} + \mathcal{X}^u_j \odot (1 - \mathcal{M}) \\ \mathcal{P}_{agg} &= \max_{k \in (1,4)} (\mathcal{F}(\mathcal{X}^u_{syn(i,j)}, [Prompt_{\#k}]; \Theta)), \end{aligned} \tag{6}$$

where $\mathcal{X}^u_{syn(i,j)}$ are mixed unlabeled data, $\mathcal{X}^u_i$ and $\mathcal{X}^u_j$ are randomly selected unlabeled data. Element-wise maximization is performed to aggregate predictions prompted by Task#1 $\sim$ Task#4, and $\mathcal{P}_{agg}$ is the final aggregated prediction. The overall loss $\mathcal{L}_{total}$ and unsupervised loss $\mathcal{L}_{unsup}$ are calculated as:

$$\begin{aligned} \mathcal{L}_{total} &= \mathcal{L}_{sup} + \mathcal{L}_{unsup} \\ \mathcal{L}_{unsup} &= Dice(\mathcal{P}_{agg}, \mathcal{F}(\mathcal{X}^u_{syn(i,j)}, [Prompt_{\#5}]; \Theta)). \end{aligned} \tag{7}$$

In summary, based on the design of semantic-aware Task#5, our VerSemi learns from unlabeled data in a task-agnostic way, and also enhances the the uniqueness of the task prompt through the auxiliary constraint $\mathcal{L}_{aux}$.

IV. Experiments

A. Datasets

Four public medical image datasets were used for this study, including **Task#1:** NIH-Pancreas [53], **Task#2:** Left Atrium [54], **Task#3:** MSD-Spleen [55] and **Task#4:** MSD-Lung Tumor [55]. The NIH-Pancreas dataset contains 82 contrast-enhanced abdominal CT scans, divided into 62 for training and 20 for testing. The Left Atrium dataset includes 100 gadolinium-enhanced MR images, with an 80/20 split for training and testing. The MSD-Spleen dataset contains 41 CT scans, split into 30 for training and 11 for testing. Lastly, the MSD-Lung Tumor dataset consists of 63 CT scans, divided into 50 for training and 13 for testing. A 10% subset of the training data is reserved for validation to determine the best model. All models adhere to this data split for fair comparisons and follow identical preprocessing protocols as described in [15] and [45].

B. Implementation Details

Following previous studies [15], [45], [47], V-Net [43] was used as the baseline model to ensure fair comparisons. The Adam optimizer [56] was employed with an initial learning rate of 0.001. The input size for the model was set to $96 \times 96 \times 96$, and the batch size was 8. All experiments were conducted using Pytorch [57] on a system equipped with four NVIDIA GeForce RTX 3080 Ti GPUs.

Each competing SSL method was trained independently for each task, in line with common practice in the field, where each task is treated in isolation. Specifically, the models were trained using the training set for each task individually, with no cross-task information sharing during the training process. After training, the models were evaluated on each task's respective test set.

C. Experiment Settings

Our proposed VerSemi framework was compared against twelve established SSL methods, including the uncertainty-aware mean-teacher (UA-MT) [44], dual-task consistency (DTC) [45], adversarial consistency and dynamic convolution (ASE-Net) [42], correlation-aware mutual learning (CAML) [46], bidirectional copy-paste [15], causality-inspired semi-supervised segmentation (CauSSL) [26], cubic volume partition and recovery (Magic-Net) [47],

context-based cross-style consistency (SLC-Net) [48], task-affinity consistency (TAC) [49], differential morphological feature perturbations (MisMatch) [50], shadow-aware network with boundary refinement (SABR-Net) [51] and virtual category learning (VC) [52].

These methods are evaluated using the Dice coefficient (%), Jaccard index (%), Average Surface Distance (ASD, in voxels), and 95% Hausdorff Distance (95HD, in voxels).

D. Results on Pancreas Dataset

Table I presents the evaluation of our VerSemi framework against twelve other SSL methods on the pancreas dataset, using 10% and 20% labeled training data. VerSemi consistently outperforms all competing methods across all metrics at both label percentages. With only 10% labeled data, VerSemi achieves a 3.07% improvement in Dice score and a 5.66-voxel reduction in 95HD compared to the second-best method, MagicNet. Additionally, VerSemi outperforms SLC-Net and VC by 4.79% and 3.32% in Dice, respectively, and reduces 95HD by 5.82 voxels and 4.31 voxels compared to SLC-Net and VC. These trends are also evident with 20% labeled data, where VerSemi continues to demonstrate superior performance, highlighting its robustness and effectiveness even with limited labeled data. Notably, VerSemi's advantages are most prominent in low-label settings, underscoring its ability to consistently outperform task-specific methods across different labeling conditions.

TABLE I

PERFORMANCE COMPARISON ON THE PANCREAS DATASET USING 10% AND 20% LABELED DATA. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND UNDERLINE, RESPECTIVELY. ((DICE, %); (JACCARD, %); (ASD, VOXEL); (95HD, VOXEL))

Method	Pancreas (10%/6 labeled data)				Pancreas (20%/12 labeled data)			
	Dice ↑	Jaccard ↑	ASD ↓	95HD ↓	Dice ↑	Jaccard ↑	ASD ↓	95HD ↓
VNet (Baseline) [43]	55.60	41.74	18.63	45.33	72.38	58.26	5.89	19.35
UA-MT (MICCAI'19) [44]	66.34	53.21	4.57	17.21	76.10	62.62	2.43	10.84
DTC (AAAI'21) [45]	69.21	54.06	5.95	17.21	78.27	64.75	2.25	8.36
ASE-Net (TMI'22) [42]	71.54	56.82	5.73	16.33	79.03	66.57	2.30	8.62
CAML (MICCAI'23) [46]	71.21	56.32	5.92	16.89	79.81	67.35	2.27	8.22
BCP (CVPR'23) [15]	73.83	59.24	3.72	12.71	82.91	70.97	2.25	6.43
CauSSL (ICCV'23) [26]	72.34	57.43	3.13	13.49	80.63	67.84	2.78	8.76
MagicNet (CVPR'23) [47]	75.01	62.04	3.97	13.71	81.25	68.81	2.83	8.50
SLC-Net (TMI'23) [48]	73.29	58.47	4.72	13.87	80.14	67.23	2.38	7.48
TAC (TMI'22) [49]	72.09	56.97	5.86	14.91	80.02	67.12	2.19	8.00
MisMatch (TMI'23) [50]	73.06	58.16	3.57	12.99	81.54	69.55	2.72	7.53
SABR-Net (TMI'23) [51]	73.65	58.81	5.16	13.95	81.47	69.03	3.27	7.91
VC (TPAMI'24) [52]	74.76	60.25	3.63	12.36	82.19	70.13	1.86	6.29
VerSemi	**78.08**	**64.82**	**2.33**	**8.05**	**83.27**	**71.68**	**1.40**	**5.33**
VerSemi w/ Task Info	78.62	64.91	2.28	7.99	83.55	71.93	1.35	5.02

TABLE II

PERFORMANCE COMPARISON ON THE LEFT ATRIUM DATASET USING 10% AND 20% LABELED DATA. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND UNDERLINE, RESPECTIVELY. ((DICE, %); (JACCARD, %); (ASD, VOXEL); (95HD, VOXEL))

Method	Left Atrium (10%/8 labeled data)				Left Atrium (20%/16 labeled data)			
	Dice ↑	Jaccard ↑	ASD ↓	95HD ↓	Dice ↑	Jaccard ↑	ASD ↓	95HD ↓
VNet (Baseline) [43]	82.74	71.72	3.26	13.35	84.89	77.32	2.97	11.60
UA-MT (MICCAI'19) [44]	86.28	76.11	4.63	18.71	88.74	79.94	2.32	8.39
DTC (AAAI'19) [45]	87.51	78.17	2.36	8.23	89.42	80.89	2.10	7.32
ASE-Net (TMI'22) [42]	87.83	78.45	2.17	9.86	90.29	82.76	1.64	7.18
CAML (MICCAI'23) [46]	**89.62**	81.28	2.02	8.76	90.78	83.19	1.68	6.11
BCP (CVPR'23) [15]	**89.62**	**81.31**	**1.76**	**6.81**	**91.25**	**83.85**	**1.47**	5.96
CauSSL (ICCV'23) [26]	88.37	79.50	2.74	9.24	90.46	82.37	1.96	6.62
MagicNet (CVPR'23) [47]	88.65	79.89	3.01	9.78	90.17	82.24	2.14	7.83
SLC-Net (TMI'23) [48]	88.79	80.19	3.45	10.47	90.15	82.23	2.32	7.33
TAC (TMI'22) [49]	87.93	79.00	3.51	11.52	90.19	82.42	2.69	6.55
MisMatch (TMI'23) [50]	88.48	79.79	3.14	11.18	90.32	82.56	2.20	7.36
SABR-Net (TMI'23) [51]	88.45	79.42	3.89	10.08	90.48	82.78	1.97	6.68
VC (TPAMI'24) [52]	88.75	80.07	3.51	9.95	90.65	83.06	2.18	8.13
VerSemi	89.01	80.52	2.57	9.03	90.89	83.48	1.72	**5.38**
VerSemi w/ Task Info	89.83	81.27	2.38	8.62	91.29	84.08	1.86	5.62

E. Results on Left Atrium Dataset

Table II presents the evaluation of our VerSemi approach, along with twelve other competing SSL methods, on the Left Atrium dataset using 10% and 20% labeled training data. VerSemi ranks second, closely following BCP with a negligible Dice difference of less than 0.5%. Additionally, VerSemi outperforms TAC, MisMatch, and SABR-Net in both Dice and Jaccard metrics across both labeling conditions. For instance, with 20% labeled data, VerSemi achieves a Dice score of 90.89, surpassing TAC, MisMatch, and SABR-Net by 0.7%, 0.57%, and 0.41%, respectively. It also reduces 95HD by at least 1.17 voxels compared to TAC and 1.30 voxels compared to SABR-Net. Notably, all methods show comparable Dice scores regardless of whether 10% or 20% of the data is labeled, suggesting minimal performance variance. This outcome is likely due to the relatively large dataset size and the increased number of slices per case, which provide sufficient labeled data for guiding the model to make accurate predictions on unlabeled data.

F. Results on Spleen Dataset

The spleen segmentation results, displayed in Table III, indicate that our VerSemi framework outperforms all other competing methods with a large margin. Notably, VerSemi surpasses MagicNet by 32.46 and 16.01 voxels in 95HD when using 10% and 20% labeled training data, respectively. Additionally, with 10% labeled data, VerSemi achieves a 7.36% improvement in Dice over CauSSL. A case study exploring the integration of data from other tasks into spleen segmentation (Task#3) further reveals consistent performance enhancements. Particularly, the integration of pancreas segmentation data (Task#1) results in notable improvements. There are three potential reasons for VerSemi's superior performance. Firstly, due to the extremely limited labels (10% corresponds to only 3 labeled cases), other methods fail to generalize the representation learned from labeled data to unlabeled data. This often results in incorrect classifying background as foreground (see bluerows 5-6 of Fig.12), leading to higher 95HD scores. Second, VerSemi can learn modality-specific

TABLE III

PERFORMANCE COMPARISON ON THE SPLEEN DATASET USING 10% AND 20% LABELED DATA. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND UNDERLINE, RESPECTIVELY. ((DICE, %); (JACCARD, %); (ASD, VOXEL); (95HD, VOXEL))

Method	Spleen (10%/3 labeled data)				Spleen (20%/6 labeled data)			
	Dice ↑	Jaccard ↑	ASD ↓	95HD ↓	Dice ↑	Jaccard ↑	ASD ↓	95HD ↓
VNet (Baseline) [43]	75.14	65.27	15.02	43.89	79.78	72.86	11.37	30.03
UA-MT (MICCAI'19) [44]	79.63	68.62	15.94	44.71	83.11	75.98	8.92	25.41
DTC (AAAI'21) [45]	80.27	69.00	14.53	41.56	84.59	75.91	9.75	31.77
ASE-Net (TMI'22) [42]	80.65	69.48	14.37	41.31	85.02	75.68	12.53	37.26
CAML (MICCAI'23) [46]	80.32	69.10	15.37	41.71	85.80	76.79	11.57	36.14
BCP (CVPR'23) [15]	83.12	72.85	14.42	42.11	87.02	78.58	10.48	37.08
CauSSL (ICCV'23) [26]	81.98	71.25	14.69	41.84	86.83	78.46	10.01	32.27
MagicNet (CVPR'23) [47]	83.55	73.58	13.49	41.79	88.24	80.24	8.50	23.51
SLC-Net (TMI'23) [48]	81.98	71.26	14.69	41.85	87.07	78.75	10.71	23.13
TAC (TMI'22) [49]	81.54	70.57	14.86	42.08	85.16	76.00	12.65	31.07
MisMatch (TMI'23) [50]	81.21	70.24	14.75	41.66	85.67	76.67	12.02	30.21
SABR-Net (TMI'23) [51]	82.25	71.59	14.26	40.87	86.21	78.22	11.18	27.04
VC (TPAMI'24) [52]	83.12	72.86	14.42	42.11	87.36	79.20	12.36	23.05
VerSemi	**89.34**	**81.73**	**3.12**	**9.33**	**94.62**	**89.89**	**2.40**	**7.50**
VerSemi w/ Task Info	90.10	82.75	3.09	9.28	94.67	89.93	2.35	7.33

TABLE IV

PERFORMANCE COMPARISON ON THE LUNG TUMOR DATASET USING 10% AND 20% LABELED DATA. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND UNDERLINE, RESPECTIVELY. ((DICE, %); (JACCARD, %); (ASD, VOXEL); (95HD, VOXEL))

Method	Lung Tumor (10%/5 labeled data)				Lung Tumor (20%/10 labeled data)			
	Dice ↑	Jaccard ↑	ASD ↓	95HD ↓	Dice ↑	Jaccard ↑	ASD ↓	95HD ↓
VNet (Baseline) [43]	31.07	21.69	14.81	24.16	36.89	26.00	12.98	23.47
UA-MT (MICCAI'19) [44]	33.46	25.88	15.08	24.78	44.33	30.80	10.28	22.82
DTC (AAAI'19) [45]	34.97	26.82	12.88	24.34	48.46	34.49	7.70	21.22
ASE-Net (TMI'22) [42]	34.18	25.86	13.09	22.91	53.15	38.29	3.77	12.87
CAML (MICCAI'23) [46]	35.24	22.99	12.33	24.25	52.43	37.03	4.07	12.65
BCP (CVPR'23) [15]	36.60	27.69	11.71	23.86	54.63	38.13	**3.62**	**11.77**
CauSSL (ICCV'23) [26]	35.72	26.25	12.52	24.09	53.69	38.47	5.05	13.40
MagicNet (CVPR'23) [47]	35.24	22.99	12.33	24.25	54.32	38.58	3.72	13.04
SLC-Net (TMI'23) [48]	35.15	22.34	14.17	25.01	53.58	39.87	5.90	12.09
TAC (TMI'22) [49]	34.84	23.61	14.38	25.31	51.69	36.90	5.72	16.87
MisMatch (TMI'23) [50]	35.14	23.32	13.72	25.35	53.46	39.89	7.74	15.92
SABR-Net (TMI'23) [51]	33.27	21.75	15.33	25.02	52.91	39.62	6.22	13.47
VC (TPAMI'24) [52]	35.29	24.04	11.70	**22.77**	54.03	40.26	5.42	13.38
VerSemi	**36.90**	**28.12**	**10.87**	23.41	**55.16**	**42.47**	6.39	16.75
VerSemi w/ Task Info	37.91	29.05	9.20	22.45	56.82	44.97	6.32	14.25

knowledge, enabling it to achieve better performance when the same imaging modality is used for Task#1 and Task#3. Furthermore, since both the pancreas (in Task#1) and spleen (in Task#3) are abdominal organs, the feature embedding of both organs are very close in the latent space (see Fig. 7). Third, VerSemi's negative learning mechanism, which involves segmenting other organs, allows it to segment background regions and identify adhesive boundaries more accurately, thereby reducing the HD score.

G. Results on Lung Tumor Dataset

The segmentation results for lung tumors, presented in Table IV, demonstrate that VerSemi achieves the highest rankings when using 10% labeled data. Specifically, VerSemi outperforms MagicNet by 5.13% in the Jaccard index. When the proportion of labeled data is increased to 20%, VerSemi maintains its lead in terms of Dice and Jaccard scores. However, it exhibits higher ASD and 95HD scores, potentially due to the misclassification of tissues surrounding the tumor that leads to the increase in surface distance metrics. Despite this, VerSemi still manages to achieve a 3.89% improvement in the Jaccard index compared to MagicNet when using 20% labeled data.

TABLE V

A CASE STUDY OF THE IMPACT OF OTHER TASKS ON ONE SPECIFIC TASK. HERE SPLEEN SEGMENTATION TASK (TASK #3) IS SELECTED AS THE BASELINE, AS WE FIND VERSEMI PRESENTS REMARKABLE IMPROVEMENTS ON THIS TASK WHEN COMPARED TO OTHER METHODS, AND THIS EXPERIMENT AIMS TO FIGURE OUT WHERE THE PERFORMANCE GAINS COME FROM

Setting	10% labels		20% labels	
	Dice ↑	95HD ↓	Dice ↑	95HD ↓
Task#3	75.14	43.89	79.78	30.03
Task#3+#1	85.62	17.07	90.00	15.81
Task#3+#1+#2	88.03	11.06	92.06	10.06
Task#3+#1+#2+#4	**89.34**	**9.33**	**94.62**	**7.50**

TABLE VI
ADAPTATION OF BCP [15] AND CAUSSL [26] TO A UNIFIED SSL MODEL, SHOWING A MARKED PERFORMANCE DECLINE IN COMPARISON TO THEIR SINGLE-TASK COUNTERPARTS

Method	Pancreas (10%/6 labeled data)				Left Atrium (10%/8 labeled data)			
	Dice ↑	Jaccard ↑	ASD ↓	95HD ↓	Dice ↑	Jaccard ↑	ASD ↓	95HD ↓
Uni-BCP	68.59	53.73	7.33	20.62	85.73	75.06	10.17	30.33
Uni-CauSSL	65.35	49.09	6.16	20.89	83.40	72.43	8.84	34.94
VerSemi w/o $\mathcal{L}_{aux}$	75.06	60.94	3.70	11.64	88.56	79.81	2.62	9.17
VerSemi (Ours)	**78.08**	**64.82**	**2.33**	**8.05**	**89.01**	**80.52**	**2.57**	**9.03**
Method	Spleen (10%/3 labeled data)				Lung Tumor (10%/5 labeled data)			
	Dice ↑	Jaccard ↑	ASD ↓	95HD ↓	Dice ↑	Jaccard ↑	ASD ↓	95HD ↓
Uni-BCP	74.80	58.89	17.11	54.06	31.01	21.32	11.35	24.36
Uni-CauSSL	73.06	57.85	18.28	55.51	25.38	20.20	15.06	28.72
VerSemi w/o $\mathcal{L}_{aux}$	87.80	79.19	3.36	10.10	34.74	22.55	12.62	24.77
VerSemi (Ours)	**89.34**	**81.73**	**3.12**	**9.33**	**36.90**	**28.12**	**10.87**	**23.41**

TABLE VII
EXPERIMENTAL RESULTS ON UNSEEN TEST SET BTCV-SPLEEN [58], ASSESSING THE GENERALIZATION ABILITY OF THE MODELS. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY

Method	**Directly testing on BTCV-spleen**				Original results on MSD-spleen (Task#3)			
	Dice ↑	Jaccard ↑	ASD↓	95HD ↓	Dice ↑	Jaccard ↑	ASD↓	95HD ↓
VNet (Baseline) [43]	73.94	61.12	14.48	43.82	75.14	65.27	15.02	43.89
MagicNet (CVPR'23) [47]	80.94	70.15	14.98	45.08	83.55	73.58	13.49	41.79
CauSSL (ICCV'23) [26]	79.34	67.78	16.89	48.02	81.98	71.25	14.69	41.84
BCP (CVPR'23) [15]	81.87	71.52	14.39	44.22	83.12	72.85	14.42	42.11
VerSemi (Ours)	89.36	83.01	2.91	10.93	89.34	81.73	3.12	9.33

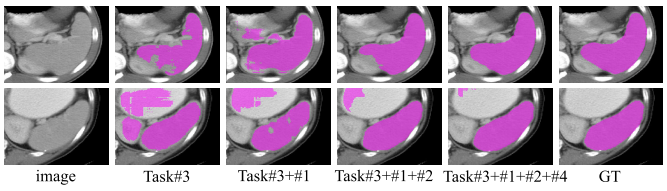

Fig. 6. Visualization of spleen segmentation results obtained by incorporating other tasks sequentially. The model tends to produce more accurate segmentation results with the increase of integrated tasks, demonstrating the benefits of learning within a unified model.

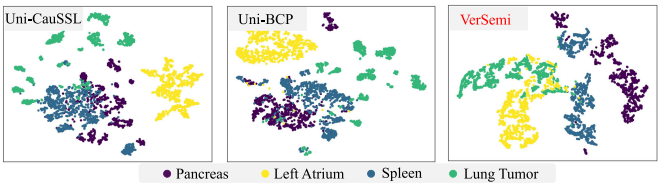

Fig. 7. t-SNE Visualization of feature embedding (from the last decoder layer of Uni-CauSSL, Uni-BCP, and our VerSemi) for four tasks.

V. DISCUSSION

A. Importance of Auxiliary Constraint $\mathcal{L}_{aux}$

The inclusion of the auxiliary constraint $\mathcal{L}_{aux}$ significantly enhances the distinctiveness of task prompts. With only 10% labeled data, integrating $\mathcal{L}_{aux}$ resulted in performance gains in the Dice scores across several segmentation tasks: 3.02% for pancreas, 0.45% for left atrium, 1.54% for spleen, and 2.16% for lung tumor segmentation (see the last two rows in Table VI). These results confirm the effectiveness of in enhancing task prompt validity, highlighting the importance of this auxiliary loss in improving segmentation performance.

B. Adapting Single SSL Models Into Unified SSL Models

We adapted CauSSL and BCP to a unified SSL framework by implementing two key modifications. First, the input data now covers four tasks, with task-specific identifiers randomly fed into the model during training to ensure task awareness. Second, the number of output channels is adjusted to match the number of tasks, differing from VerSemi, which uses a dynamic task-prompted head with only two output channels. These adjustments allow for a more equitable comparison between the models under a joint training paradigm. The results presented in Table VI demonstrate that both Uni-BCP and Uni-CauSSL underperform VerSemi. Moreover, when these unified models are compared to their original single-task versions, we observe a significant decline in performance. For example, BCP versus Uni-BCP shows a 5.76% decrease in average Dice score with 10% labeled data (from 70.79% to 65.03%), and CauSSL versus Uni-CauSSL shows a similar 7.80% reduction (from 69.60% to 61.80%). These results suggest that a naive approach to multi-task learning, without careful consideration of task-specific challenges, can undermine the performance of individual tasks. Additionally, t-SNE visualizations of feature embeddings, shown in Fig. 7, further highlight the advantage of VerSemi. While VerSemi maintains clear decision boundaries between tasks, other models exhibit overlapping and dispersed feature embeddings. This underscores the critical role of task prompts and the auxiliary constraint ($\mathcal{L}_{aux}$) in managing multiple SSL tasks. Task prompts guide the model in focusing on the current task, while

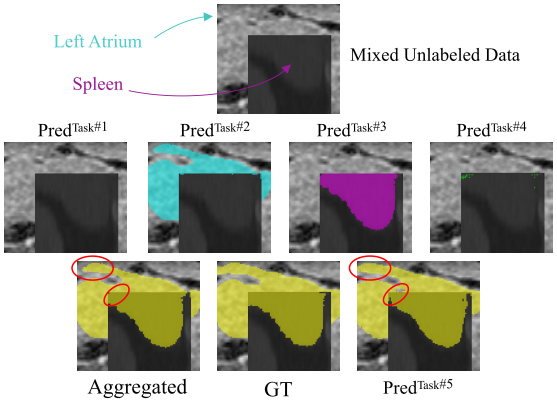

Fig. 8. Visualization of segmentation results prompted by relevant tasks and synthetic Task#5. The spleen and left atrium are mixed to generate mixed unlabeled data. The inconsistent regions between aggregated prediction and Task#5-prompted prediction are highlighted by red elliptic.

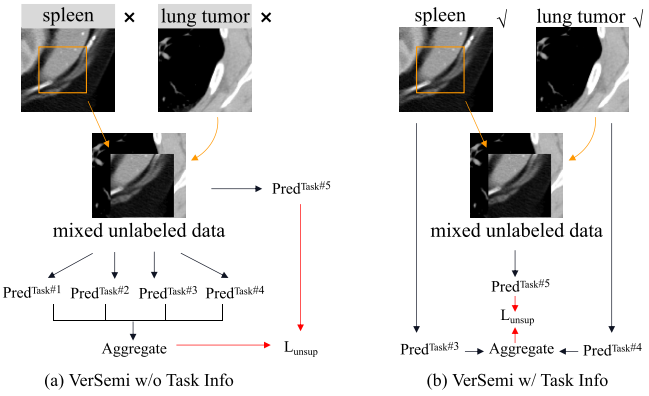

Fig. 9. (a) VerSemi is designed to learn from unlabeled data without task-specific information. In this case, VerSemi is prompted by all related tasks (Task#1 ∼ Task#4) on mixed images. This operation aims to obtain task-specific predictions that are then aggregated and aligned with Task#5 prediction. (b) The pipeline of VerSemi when unlabeled task information is available, allowing direct predictions on source images with task-specific prompts.

the auxiliary constraint ensures that the learned representations for each task remain distinct and concentrated.

C. Visualization of Unlabeled Data Learning Pipeline

Fig. 8 shows the segmentation results when prompted by relevant tasks and the synthetic Task#5. It reveals that VerSemi can accurately recognize regions associated with specific task prompts, greatly aided by the auxiliary constraint $\mathcal{L}_{aux}$, which enhances the controllability of prompts. Meanwhile, VerSemi adeptly delineates all task semantic regions under the prompt of Task#5, showcasing the model's adeptness at learning a semantic-aware synthetic task. By aligning the two predictions (see the bottom of Fig. 8, Aggregated and Perd$^{Task\#5}$), it is evident that VerSemi is capable of learning from unlabeled data in a task-agnostic fashion.

D. Incorporating Unlabeled Task Information Into VerSemi

We further investigated the potential of VerSemi by integrating the task information from unlabeled data. As shown in the results from Table I to Table IV, VerSemi w/ Task Info outperforms the standard VerSemi by a 0.78% improvement in the average Dice score with 10% labeled data. Fig. 9 highlights the key differences between the two approaches. In VerSemi, predictions are aggregated from multiple tasks (Task#1 to Task#4) and aligned with the synthetic predictions from Task#5. In contrast, VerSemi w/ Task Info is aware of the specific task associated with the unlabeled data, enabling it to make direct predictions on the source images before aligning them with Task#5's predictions. The main distinction between VerSemi and VerSemi w/ Task Info lies in how the aggregated predictions are generated. In VerSemi, the predictions are based solely on the mixed data, which introduces challenges, even when task-specific information is used for prompting. On the other hand, VerSemi w/ Task Info generates predictions from source images, unaffected by data from other tasks. This results in higher-quality aggregated predictions, leading to improved overall performance in VerSemi w/ Task Info compared to VerSemi.

TABLE VIII
COMPARISON OF THREE TYPES OF PROMPT: LANGUAGE, SOFT VECTOR, AND ONE-HOT VECTOR PROMPTS. THE AVERAGED DICE AND 95HD SCORES ON FOUR TASKS ARE REPORTED

Types of prompt	10% labels		20% labels	
	Dice ↑	95HD ↓	Dice ↑	95HD ↓
language (CLIP) [41]	67.21	22.78	74.30	19.29
soft vector	70.02	17.66	77.45	13.83
one-hot vector	**73.33**	**12.46**	**80.99**	**8.74**

E. Discussion of Task Prompt

Task prompts serve as cues to assist the model in identifying the current task. Common prompt formats include language descriptions [24], [59] (using natural language sentences to describe tasks), soft vector prompts [21], [60] (using randomly initialized learnable vectors) and one-hot prompts [6], [61]. As shown in Table VIII, the one-hot prompt performs best in an SSL context. This can be attributed to two main factors: (1) language embeddings, which rely on pre-trained language or vision-language models, may not align well with the embeddings used for medical images, and (2) soft vector prompts require large amounts of paired image-label data, which are scarce in SSL settings. In contrast, the explicit nature of one-hot prompt makes it more suitable and effective for SSL tasks, as it does not depend on external embeddings or large labeled datasets.

F. Learned Distribution on Labeled and Unlabeled Data

Discrepancies between the distributions of labeled and unlabeled data pose a common challenge in SSL, often due to an unbalanced or partial distribution learned from the labeled data [7], [47]. Fig. 10 illustrates the kernel density estimation for VerSemi and BCP trained with a 10% label ratio. The findings indicate that for Task#2 (left atrium segmentation), which has a larger dataset, both models exhibit well-aligned distributions, likely due to the comprehensive representation learned from the labeled data, allowing the

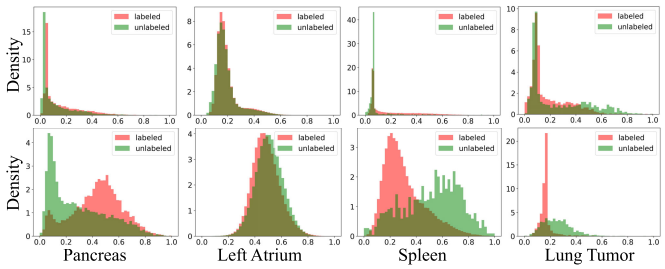

Fig. 10. Kernel density estimation of (top) VerSemi and (bottom) BCP [15] when trained with 10% labels, illustrating VerSemi's superior alignment of labeled and unlabeled data distributions compared to BCP.

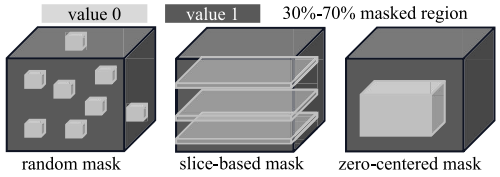

Fig. 11. Discussion of three types of copy-past choices. The random mask involves cropping several random cubic patches of size 6 × 6×6. The slice-based mask generates slice-wise masks, effectively fusing different unlabeled data at the slice level. The zero-centered mask involves cropping larger patches, where the proportion of the masked area is randomly sampled between 0.3 and 0.7.

models to generalize effectively to unlabeled data. However, for the other tasks, BCP shows severe inconsistencies, while VerSemi maintains a more aligned distribution. The results suggest that properly learning multiple tasks concurrently is beneficial for unlabeled data mining, since the mismatch issue between labeled and unlabeled data is largely alleviated.

G. Generalizing to Unseen Test Set

To evaluate the generalization ability of SSL models, an experiment was conducted using BTCV-spleen as an unseen test set, consisting of 30 samples. The models, trained with 10% labeled data, were tested directly without fine-tuning. The original results for MSD-spleen (Task#3) are presented in the right part of Table VII, while the opposite side presents the results on all metrics on the unseen test set. VerSemi consistently outperforms other models across all metrics on the unseen test set. While other SSL methods exhibit a roughly 2% decline in Dice score when comparing BTCV-spleen to MSD-spleen results, VerSemi shows virtually no decrease, illustrating its robustness in handling data from unknown sources and also affirming the superiority of our proposed paradigm of integrating various related SSL tasks into a unified framework.

H. Discussion of Copy-Past Choices

Task#5 is constructed using the CutMix technique, with variations in the cropping strategy for the mask generation. As shown in Table IX, three cropping strategies are evaluated: the random mask, slice-based mask, and zero-centered mask (see Fig.11). The random mask involves cropping several random cubic patches of size 6 × 6×6. The slice-based mask creates slice-wise masks, effectively merging different unlabeled data at the slice level. The zero-centered mask crops larger patches, with the masked area proportion randomly sampled between 0.3 and 0.7. The results indicate that the random mask yields the poorest performance. The randomly generated cubic patches lack coherent foreground information, making them more likely to be confused with the background of other tasks. In contrast, the slice-based mask improves performance by preserving the integrity of the foreground. The zero-centered mask outperforms both, as it facilitates interaction between the foreground and background across tasks while maintaining target coherence. Therefore, based on these findings, CutMix with the zero-centered mask is used in the construction of Task#5.

TABLE IX
DESIGN CHOICES OF COPY-PAST FOR THE GENERATION OF TASK#5. THE AVERAGED DICE AND 95HD SCORES ON FOUR TASKS ARE REPORTED UNDER LABEL PERCENTAGE OF 10% AND 20%

Setting	10% labels		20% labels	
	Dice ↑	95HD ↓	Dice ↑	95HD ↓
Random mask	70.21	15.32	78.79	11.71
Slice-based mask	72.16	13.88	79.27	10.55
Zero-centered mask	**73.33**	**12.46**	**80.99**	**8.74**

TABLE X
EFFICIENCY ANALYSIS WITH AN INPUT SIZE OF 96 × 96×96. METRICS INCLUDE MULTIPLY-ACCUMULATE OPERATIONS (MACS), TRAINING TIME, INFERENCE TIME, AND PARAMETERS (PARA.) OF THE MODEL

Method	Para. (M)	MACs (G)	Training	Test / case
CauSSL [26]	18.90	83.88	∼ 17h36mins	∼ 2.96sec
VerSemi (Ours)	9.48	35.50	∼ 19h12mins	∼ 2.81sec

I. Computational Complexity

Table X presents an efficiency analysis using metrics such as multiply-accumulate operations (MACs), parameters, training time, and test time. The results indicate that VerSemi has lower parameters, MACs, and test time compared to the competing CauSSL [26], underscoring VerSemi's superiority in integrating multiple tasks while maintaining efficiency. Specifically, CauSSL utilizes two independent networks in a co-training approach with causal mechanisms. In contrast, our VerSemi jointly learns from multiple semi-supervised tasks using a single network, where the parameters of the segmentation head are dynamically generated with task-specific prompts, resulting in faster inference times for VerSemi. Due to the over-sampling strategy employed in VerSemi to mitigate the influence of varying data sizes across different datasets, the total training time on four datasets is higher than that of CauSSL. However, this approach enhances overall performance compared to CauSSL, which is trained on each dataset in isolation.

J. Visualization of Segmentation Results

We visulize the segmentation results on four benchmarks in Fig. 12. It indicates that VerSemi produces the most precise segmentation masks compared to the competing methods. For

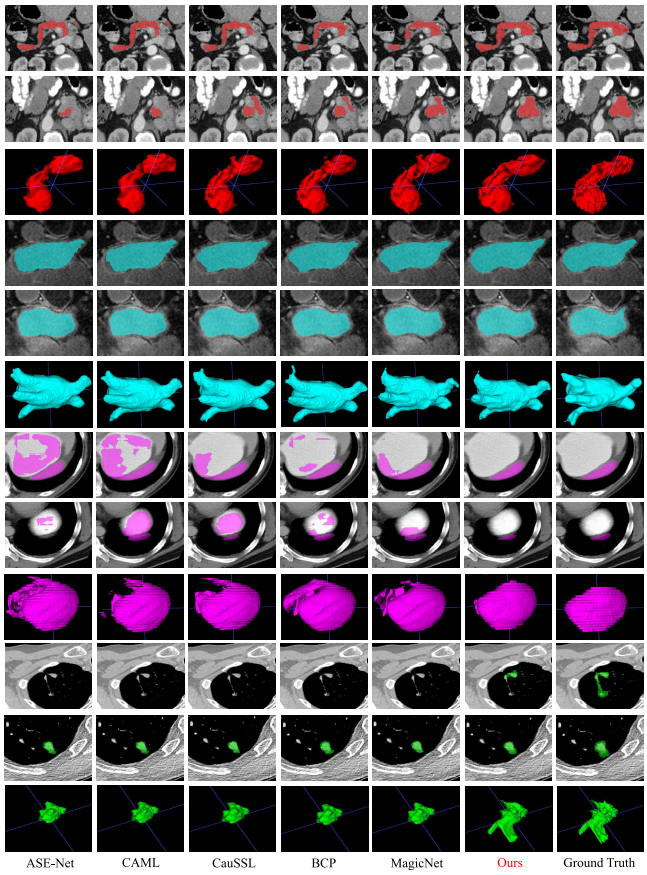

Fig. 12. Segmentation results produced by different methods across four tasks, with rows 1-3 for pancreas, 4-6 for left atrium, 7-9 for spleen, and 10-12 for lung tumor.

instance, in spleen segmentation (Rows 5-6), while other SSL methods extensively misclassify background as foreground, VerSemi accurately identifies the spleen region. In the case of left atrium segmentation (Rows 2-3), most SSL methods achieve near-perfect results, likely due to the larger dataset size, which provides sufficient labeled data for the models to generalize effectively to the test set.

VI. Conclusion

This paper presents VerSemi, an effective model for semi-supervised medical image segmentation that innovatively integrates various tasks into a unified framework. VerSemi dynamically handles different tasks through the design of task prompts, and a novel contrastive constraint is proposed to improve the controllability of these dynamic task prompts, thereby distinguishing between different task information. Extensive experiments on four public datasets have demonstrated the effectiveness of our VerSemi model, particularly with limited training labels, setting a new state of the art for semi-supervised medical image segmentation. Furthermore, VerSemi's integration of multiple tasks into a unified framework enables it to generalize across a variety of clinical tasks, such as the segmentation of organs, tumors, and other anatomical structures. The framework also exhibits flexibility across different imaging modalities, including CT and MRI, making it highly adaptable to a wide range of medical imaging applications. This broad applicability ensures that VerSemi can be deployed in diverse healthcare settings, enhancing the consistency and reliability of medical image analysis on a global scale. Additionally, VerSemi's ability to effectively handle limited annotated data addresses a significant challenge in real-world clinical environments, where expert annotations are often scarce or costly. By leveraging large volumes of unlabeled data, VerSemi reduces the reliance on manual labeling, thereby accelerating workflows and making automated segmentation tools more accessible, particularly in resource-constrained settings. **Limitation and Future Work.** Since our model has been trained on several available datasets, the inherent inter-dataset conflicts may impact the training process. Future work will focus on developing a de-biased strategy to handle such inter-dataset conflicts.

References

[1] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, no. 1, pp. 221–248, 2017.
[2] V. Cheplygina, M. de Bruijne, and J. P. W. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Med. Image Anal.*, vol. 54, pp. 280–296, May 2019.
[3] R. Jiao, Y. Zhang, L. Ding, R. Cai, and J. Zhang, "Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation," 2022, *arXiv:2207.14191*.
[4] Y. Wu et al., "Diversified and personalized multi-rater medical image segmentation," in *Proc. CVPR*, 2024, pp. 11470–11479.
[5] Y. Wu, Z. Wu, H. Shi, B. Picker, W. Chong, and J. Cai, "CoactSeg: Learning from heterogeneous data for new multiple sclerosis lesion segmentation," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2023, pp. 3–13.
[6] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "DoDNet: Learning to segment multi-organ and tumors from multiple partially labeled datasets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1195–1204.
[7] Q. Zeng, Y. Xie, Z. Lu, and Y. Xia, "PEFAT: Boosting semi-supervised medical image classification via pseudo-loss estimation and feature adversarial training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15671–15680.
[8] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, Feb. 2020.
[9] Q. Zeng, J. Geng, K. Huang, W. Jiang, and J. Guo, "Prototype calibration with feature generation for few-shot remote sensing image scene classification," *Remote Sens.*, vol. 13, no. 14, p. 2728, Jul. 2021.
[10] W. Zhang et al., "BoostMIS: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation," in *Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 20666–20676.
[11] K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 596–608.
[12] Q. Zeng, J. Geng, W. Jiang, K. Huang, and Z. Wang, "IDLN: Iterative distribution learning network for few-shot remote sensing image scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.
[13] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2018.
[14] Q. Zeng, Y. Xie, Z. Lu, and Y. Xia, "A human-in-the-loop method for pulmonary nodule detection in CT scans," *Vis. Intell.*, vol. 2, no. 1, p. 19, Jul. 2024.
[15] Y. Bai, D. Chen, Q. Li, W. Shen, and Y. Wang, "Bidirectional copy-paste for semi-supervised medical image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11514–11524.
[16] X. Ma et al., "Point set registration with mixture framework and variational inference," *Pattern Recognit.*, vol. 104, Aug. 2020, Art. no. 107345.

[17] Z. Wang, C. Liu, S. Zhang, and Q. Dou, "Foundation model for endoscopy video analysis via large-scale self-supervised pre-train," 2023, *arXiv:2306.16741*.

[18] Y. Xie, J. Zhang, Y. Xia, and Q. Wu, "UniMiSS: Universal medical self-supervised learning via breaking dimensionality barrier," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2022, pp. 558–575.

[19] H.-Y. Zhou, C. Lian, L. Wang, and Y. Yu, "Advancing radiograph representation learning with masked record modeling," in *Proc. ICLR*, 2022.

[20] J. Chen et al., "CancerUniT: Towards a single unified model for effective detection, segmentation, and diagnosis of eight major cancers using a large collection of CT scans," 2023, *arXiv:2301.12291*.

[21] Y. Ye, Y. Xie, J. Zhang, Z. Chen, and Y. Xia, "UniSeg: A prompt-driven universal segmentation model as well as a strong representation learner," 2023, *arXiv:2304.03493*.

[22] A. Kirillov et al., "Segment anything," 2023, *arXiv:2304.02643*.

[23] L. Xue et al., "ULIP: Learning a unified representation of language, images, and point clouds for 3D understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1179–1189.

[24] J. Liu et al., "Clip-driven universal model for organ segmentation and tumor detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2023, pp. 21152–21164.

[25] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6023–6032.

[26] J. Miao, C. Chen, F. Liu, H. Wei, and P.-A. Heng, "CauSSL: Causality-inspired semi-supervised learning for medical image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 21369–21380.

[27] Y. Chen, M. Mancini, X. Zhu, and Z. Akata, "Semi-supervised and unsupervised deep visual learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 3, pp. 1327–1347, Mar. 2024.

[28] A. Mey and M. Loog, "Improved generalization in semi-supervised learning: A survey of theoretical results," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4747–4767, Apr. 2023.

[29] Y. Wu et al., "Mutual consistency learning for semi-supervised medical image segmentation," *Med. Image Anal.*, vol. 81, Oct. 2022, Art. no. 102530.

[30] Q. Zeng and J. Geng, "Task-specific contrastive learning for few-shot remote sensing image scene classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 191, pp. 143–154, Sep. 2022.

[31] Q. Zeng, Y. Xie, Z. Lu, M. Lu, J. Zhang, and Y. Xia, "Consistency-guided differential decoding for enhancing semi-supervised medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 44, no. 1, pp. 44–56, Jan. 2025.

[32] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," in *Proc. ICLR*, 2020.

[33] H. Chen et al., "SoftMatch: Addressing the quantity-quality tradeoff in semi-supervised learning," in *Proc. ICLR*, 2022.

[34] Y. Wang et al., "FreeMatch: Self-adaptive thresholding for semi-supervised learning," in *Proc. ICLR*, 2022.

[35] D. Kim, J. Kim, S. Cho, C. Luo, and S. Hong, "Universal few-shot learning of dense prediction tasks with visual token matching," in *Proc. ICLR*, 2022.

[36] H. Lee et al., "Vision-language generative model for view-specific chest X-ray generation," 2023, *arXiv:2302.12172*.

[37] Q. Zeng, Z. Lu, Y. Xie, M. Lu, X. Ma, and Y. Xia, "Reciprocal collaboration for semi-supervised medical image classification," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2024, pp. 522–532.

[38] Z. Lu, Y. Xie, Q. Zeng, M. Lu, Q. Wu, and Y. Xia, "Spot the difference: Difference visual question answering with residual alignment," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Jan. 2024, pp. 649–658.

[39] X. Ma, H. Cui, S. Li, Y. Yang, and Y. Xia, "Deformable medical image registration with global–local transformation network and region similarity constraint," *Computerized Med. Imag. Graph.*, vol. 108, Sep. 2023, Art. no. 102263.

[40] Q. Zeng, Z. Lu, Y. Xie, and Y. Xia, "PICK: Predict and mask for semi-supervised medical image segmentation," *Int. J. Comput. Vis.*, pp. 1–16, 2025, doi: 10.1007/s11263-024-02328-9.

[41] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[42] T. Lei, D. Zhang, X. Du, X. Wang, Y. Wan, and A. K. Nandi, "Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network," *IEEE Trans. Med. Imag.*, vol. 42, no. 5, pp. 1265–1277, May 2023.

[43] F. Milletari, N. Navab, and S. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[44] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 605–613.

[45] X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 8801–8809.

[46] S. Gao, Z. Zhang, J. Ma, Z. Li, and S. Zhang, "Correlation-aware mutual learning for semi-supervised medical image segmentation," in *Proc. MICCAI*, Jan. 2023, pp. 98–108.

[47] D. Chen, Y. Bai, W. Shen, Q. Li, L. Yu, and Y. Wang, "MagicNet: Semi-supervised multi-organ segmentation via magic-cube partition and recovery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23869–23878.

[48] J. Liu, C. Desrosiers, D. Yu, and Y. Zhou, "Semi-supervised medical image segmentation using cross-style consistency with shape-aware and local context constraints," *IEEE Trans. Med. Imag.*, vol. 43, no. 4, pp. 1449–1461, Apr. 2024.

[49] J. Chen, J. Zhang, K. Debattista, and J. Han, "Semi-supervised unpaired medical image segmentation through task-affinity consistency," *IEEE Trans. Med. Imag.*, vol. 42, no. 3, pp. 594–605, Mar. 2023.

[50] M.-C. Xu et al., "MisMatch: Calibrated segmentation via consistency on differential morphological feature perturbations with limited labels," *IEEE Trans. Med. Imag.*, vol. 42, no. 10, pp. 2988–2999, Oct. 2023.

[51] F. Chen et al., "Deep semi-supervised ultrasound image segmentation by using a shadow aware network with boundary refinement," *IEEE Trans. Med. Imag.*, vol. 42, no. 12, pp. 3779–3793, Dec. 2023.

[52] C. Chen, J. Han, and K. Debattista, "Virtual category learning: A semi-supervised learning method for dense prediction with extremely limited labels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5595–5611, Aug. 2024.

[53] H. R. Roth et al., "DeepOrgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 556–564.

[54] Z. Xiong et al., "A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101832.

[55] M. Antonelli et al., "The medical segmentation decathlon," *Nat. Commun.*, vol. 13, no. 1, p. 4128, 2022.

[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[57] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, vol. 32, 2019, pp. 8024–8035.

[58] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein, "MICCAI multi-atlas labeling beyond the cranial vault–workshop and challenge," in *Proc. MICCAIW*, vol. 5, 2015, p. 12.

[59] H. Yao, R. Zhang, and C. Xu, "Visual-language prompt tuning with knowledge-guided context optimization," in *Proc. CVPR*, 2023, pp. 6757–6767.

[60] T. Vu, B. Lester, N. Constant, R. Al-Rfou, and D. Cer, "SPoT: Better frozen model adaptation through soft prompt transfer," 2021, *arXiv:2110.07904*.

[61] M. Yasunaga, J. Leskovec, and P. Liang, "LinkBERT: Pretraining language models with document links," 2022, *arXiv:2203.15827*.